A5

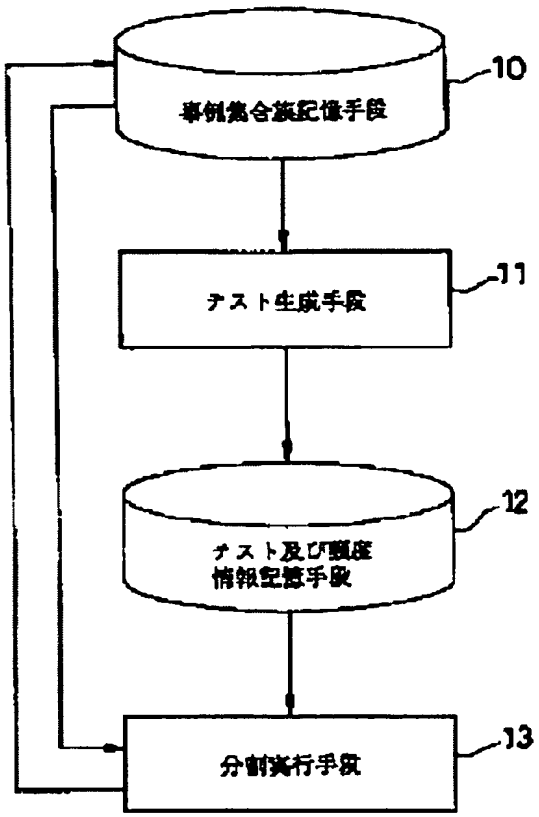
DETERMINATION TREE GENERATING METHOD

Patent number: JP9305404
Publication date: 1997-11-28
Inventor: AKIBA YASUHIRO
Applicant: NIPPON TELEGR & TELEPH CORP <NTT>
Classification:
- international: G06F9/44; G06F9/44
- european:
Application number: JP19960117890 19960513
Priority number(s):

Abstract of **JP9305404**

PROBLEM TO BE SOLVED: To simply express a determination tree by selecting a question, that has the most optimum training cases to be covered by the question, as a question to an attribute when there are plural optimum questions to the respective attributes having tree structure.

SOLUTION: When there are plural optimum questions to the respective attributes having the tree structure, the question having the most training case to be covered by that question is generated as the question to the attribute by a test generating means 11. That generated test is stored in a test and frequency information storage means 12, that stored test is supplied to a divided executing means 13, that divided executing means 13 selects any optimum test out of the test supplied from the test and frequency information storage means 12 and a test in the manner of ID 3, divides the test into mutually primary case sets and stores them in a case set storage means 10. By such a method, the determination tree can be simply expressed.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平9－305404

(43)公開日　平成９年(1997)11月28日

(51)Int.Cl.⁶	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　9/44	５５０		Ｇ０６Ｆ　9/44	５５０Ｎ	
	５８０			５８０Ｅ	

審査請求　未請求　請求項の数２　ＯＬ　（全　６　頁）

(21)出願番号　特願平8－117890

(22)出願日　平成８年(1996)５月13日

(71)出願人　000004226
日本電信電話株式会社
東京都新宿区西新宿三丁目19番２号

(72)発明者　秋葉　泰弘
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

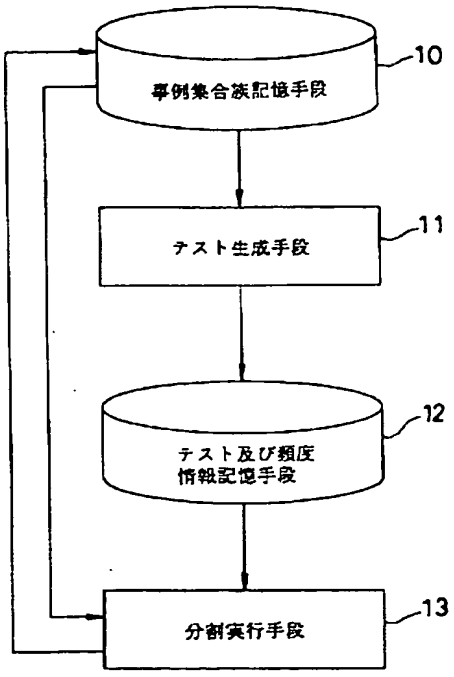
(74)代理人　弁理士　三好　秀和　（外１名）

(54)【発明の名称】　決定木生成法

(57)【要約】

【課題】　生成される決定木が木構造を有する属性で表現される場合には、該木構造上でのより上位のノードに対する質問が選択される決定木生成法を提供することにある。

【解決手段】　訓練事例の表現に利用する属性の少なくとも１つが木構造を有する時、木構造を有する各属性に対して最適な質問が複数ある場合、該質問によってカバーされる訓練事例が一番多く質問を該属性に対する質問として選択し、これらの事例を説明する所望の決定木を生成する。



1

【特許請求の範囲】

【請求項１】　いくつかの属性値と該属性値から決定されるクラスで表現された判断事例集合からルールを生成する決定木生成法であって、訓練事例の表現に利用する属性の少なくとも１つが木構造を有する時、木構造を有する各属性に対する最適な質問が複数ある場合、その質問によってカバーされる訓練事例が一番多い質問を該属性に対する質問として選択することを特徴とする決定木生成法。

【請求項２】　いくつかの属性値と該属性値から決定されるクラスで表現された判断事例集合からルールを生成する決定木生成法であって、訓練事例の表現に利用する属性の少なくとも１つが木構造を有する時、木構造を有する各属性に対する最適な質問が複数ある場合、各質問の評価値を計算し、この計算した各質問の評価値のうち最も大きな評価値を有する質問を該属性に対する質問として選択することを特徴とする決定木生成法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、属性値とクラスで表現される判断事例集合からルールを生成する決定木生成法に関し、更に詳しくは、専門家の経験をルールとして計算機の中に取り込んで、専門家と同等の判断を可能とすることを特徴とするエキスパートシステムの構築に必要なルールを専門家の判断事例から自動的に生成する決定木生成法に関する。

【０００２】

【従来の技術】近年、専門家の経験をルールとして、計算機の中に取り込んで、専門家と同等の判断をすることを特徴とするエキスパートシステムの開発・研究が活発である。しかし、エキスパートシステム構築の大きな問題点に、専門家の知識をルールとして整理することの困難さが上げられている。この問題に対処するため、専門家の判断事例から自動的にルールを生成する方法の研究が盛んである。

【０００３】専門家の判断事例から自動的にルールを生成する代表的な方法の１つに、オーストラリアのQuinlan により開発されたＩＤ３がある（「知識獲得入門・帰納学習と応用」、第５章帰納的推論機構、共立出版株式会社発行参照）。ＩＤ３による事例からのルール獲得では、事例は、ある決まった属性ベクトルにより記述される。属性ベクトルの各成分の取り得る値は、属性の取り得る値である。その例を示せば、「色」とか「ボーレート」などの属性の取り得る値、｛赤、緑、青｝、あるいは｛１００，３００，１２００，４８００｝が各成分の取り得る値である。ＩＤ３では、ルールは、決定木として獲得される。

【０００４】事例集合Ｅに対して、事例集合Ｅの要素が同一のクラスに属している時の決定木は、そのクラスの名前を持つ１つの葉そのものである。また、事例集合Ｅ

2

を２つ以上のクラスにクラス分けできるときには、適当な属性を選択することにより集合Ｅをそれぞれ互いに素な部分集合Ｅ１，Ｅ２，Ｅ３，…，Ｅｎ に分割することができる。ここで、Ｅｉ はＥの要素の中で選択された属性のｉ番目の属性値を成分に持つ要素を含む部分集合である。更に、これらの部分集合のそれぞれは、いま述べたような生成規則手順を繰り返すことによって更に処理することができる。最終的に結果は木構造となり、それぞれの葉部分にはクラスの名前が付けられている。それまでの各ノードにおいては、テストされるべき属性を規定しており、そこからの枝別れは、その属性が得られる値に対応している。

【０００５】上述した過程を説明するために、次の集合を考える。対象は、３つの属性「背丈」、「髪の色」、「目の色」により、２つのクラス「＋」、「－」に対応づけられている。

Ｅ　＝　低い、ブロンド、青；＋
低い、黒、青；－
高い、黒、茶色；－
高い、ブロンド、茶；－
高い、黒、青；－
低い、ブロンド、茶；－
高い、赤、青；＋
高い、ブロンド、青；＋

【０００６】すべての事例が同一のクラスに属しているわけではないので、テストが必要である。テストは、３つの属性「背丈」、「髪の色」、「目の色」に対して可能である。

【０００７】例えば、髪の色に対してテストを行うと、判断木は、図３に示すようになる。髪の色が「黒」と「赤」の集合については、単一のクラスしか含まれていないので、これ以上分類を行う必要はない。

【０００８】一方、「ブロンド」のクラスについては、複数のクラスが存在するため、同様の手順を繰り返す必要がある。

【０００９】上記の分類では、「髪の色」を分類に用いた。しかし、「背丈」、「目の色」に対しても、分類は可能であり、３つの属性の中でどれを用いるかが実際には問題となる。ＩＤ３では、この問題に対処するため、３つの属性の各々を用いて分類を行った場合の分類結果のエントロピーを計算し、最もエントロピーの減少の大きな属性を用いて分類を行うことを提案している。このＩＤ３は、最小の分類回数で分類を終了するものではないが、ある事例を分類してクラスを決定する際に、その判断回数が（準）最小となる実用的な方法として、多用されている。上記事例集合Ｅに対する最終的な決定木は、図４に示すようなものである。

【００１０】事例を表現する属性の中には、そこに入りえる属性値が、意味的な上下関係をもつため、木構造を有するものがある。例えば、事例を表現する属性の中

に、属性“形”があり、属性値が、｛三角形、六角形、正方形、真の楕円、円、凹｝であったとすると、属性“形”は、図５に示すような木構造をもつ。このような事例を表現する属性の中に木構造を有する場合に、自動的に決定木を生成する代表的な方法の１つにAlmuallimらによる方法がある（「On handling Tree-Structured Attributes in Decision Tree learning」in Proceedings of the 12th International Conference, Morgan Kaufmann 発行参照）。

【００１１】このAlmuallim らによる方法は、木構造を有する属性について、対応する木構造Ｔに訓練事例を流し、訓練事例が通った木構造Ｔの任意のノードＮに対して、“訓練事例の該属性の属性値が、木構造Ｔ上でノードＮの下位ノードであるか”という質問を生成する。例えば、対象が、４つの属性「背丈」、「髪の色」、「目の色」、「好きな形」により、２つのクラスに対応づけられていて、「好きな形」が図５の木構造Ｔをもっている次の事例集合：

S1 ＝ 低い、ブロンド、青、三角形；＋
低い、黒、青、六角形；－
高い、黒、茶、凹；－
高い、ブロンド、茶、正方形；－
高い、黒、青、凹；－
低い、ブロンド、茶、凹；－
高い、赤、青、三角形；＋
高い、ブロンド、青、正方形；＋

を考えると、木構造をもつ属性は「好きな形」だけで、この属性に対して以下の６つのテスト：

（１）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「凸」の下位ノードであるか？

（２）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「正多角形」の下位ノードであるか？

（３）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「三角形」の下位ノードであるか？

（４）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「六角形」の下位ノードであるか？

（５）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「正方形」の下位ノードであるか？

（６）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「凹」の下位ノードであるか？

を生成する。最終的に、それらの質問の中で、最もエントロピーの減少の大きな質問を該属性に対する質問候補としている。

【００１２】

【発明が解決しようとする課題】上記の決定木を生成する従来の技術には、大きな欠点がある。従来法により生成された決定木が木構造Ｔを有する属性で表現さる場合には、該木構造Ｔの上でより上位のノードに対する質問が選択された方が、決定木が簡潔に表現され、同業者の間で広く知られているオッカムの剃刀の原理では、好ましい。しかし、従来法では、必ずしも上位のノードに対する質問が選択されるとは限らない。

【００１３】本発明は、上記に鑑みてなされたものであり、その目的とするところは、生成される決定木が木構造Ｔを有する属性で表現される場合には、該木構造Ｔの上でより上位のノードに対する質問が選択される決定木生成方法を提供することにある。

【００１４】

【課題を解決するための手段】上記目的を達成するため、請求項１記載の本発明は、いくつかの属性値と該属性値から決定されるクラスで表現された判断事例集合からルールを生成する決定木生成法であって、訓練事例の表現に利用する属性の少なくとも１つが木構造を有する時、木構造を有する各属性に対する最適な質問が複数ある場合、その質問によってカバーされる訓練事例が一番多い質問を該属性に対する質問として選択することを要旨とする。

【００１５】請求項１記載の本発明にあっては、訓練事例の表現に利用する属性の少なくとも１つが木構造を有する時、木構造を有する各属性に対して最適な質問が複数ある場合、該質問によってカバーされる訓練事例が一番多く質問を該属性に対する質問として選択し、これらの事例を説明する所望の決定木を生成する。

【００１６】また、請求項２記載の本発明は、いくつかの属性値と該属性値から決定されるクラスで表現された判断事例集合からルールを生成する決定木生成法であって、訓練事例の表現に利用する属性の少なくとも１つが木構造を有する時、木構造を有する各属性に対する最適な質問が複数ある場合、各質問の評価値を計算し、この計算した各質問の評価値のうち最も大きな評価値を有する質問を該属性に対する質問として選択することを要旨とする。

【００１７】請求項２記載の本発明にあっては、訓練事例の表現に利用する属性の少なくとも１つが木構造を有する時、木構造を有する各属性に対して最適な質問が複数ある場合、各質問の評価値を計算し、この計算した各質問の評価値のうち最も大きな評価値を有する質問を該属性に対する質問として選択し、これらの事例を説明する所望の決定木を生成する。

【００１８】

【発明の実施の形態】以下、図面を用いて本発明の実施の形態について説明する。

【００１９】図１は、本発明の一実施形態に係る決定木生成法を実施する装置の構成を示すブロック図である。同図に示す装置は、生成される決定木が木構造Ｔを有する属性で表現さる場合には、該木構造Ｔの上、より上位のノードに対する質問が選択された決定木を生成するものであり、処理の途中の分割された事例集合の集合である事例集合族を記憶している事例集合族記憶手段１０を有する。

5

【００２０】この事例集合族記憶手段１０に記憶されている事例集合族の各々の集合には、各事例を表現する複数の属性値と該事例が属するクラスとが記憶されている。事例集合族記憶手段１０に記憶されている事例集合族は、テスト生成手段１１に供給される。テスト生成手段１１は、木構造を有する各属性Ａについて、最適な質問が複数あれば、その質問によってカバーされる訓練事例が一番多い質問を該属性に対する質問として生成する。テスト生成手段１１で生成されたテストは、テスト及び頻度情報記憶手段１２に記憶される。テスト及び頻度情報記憶手段１２に記憶されたテストは、分割実行手段１３に供給され、分割実行手段１３はテスト及び頻度情報記憶手段１２から供給されたテストおよびＩＤ３流のテストから最適なテストを選択し、互いに素な事例集合を分割し、事例集合族記憶手段１０に記憶する。いま述べたように生成規則手順を繰り返すことによって、最終的に結果は木構造となり、それぞれの葉部分にはクラスの名前が付けられている。

【００２１】次に、図２を参照して、テスト生成手段１１の処理内容の実現方法を更に詳しく説明する。テスト生成手段１１は、事例集合族記憶手段１０に記憶されている事例集合族から、テストを生成するものである。木構造をもつ各属性毎に図２に示す処理を行う。

【００２２】まず、処理を開始すると、木構造を有する属性Ａには、Ｎ個の質問候補Ｕｉ （ｉ＝１，２，…，Ｎ）が可能であるとし、最適な質問番号格納領域Ｊとカウンターｊをテスト及び頻度情報記憶手段１２に準備し、また各質問候補Ｕｉ の評価値を格納する評価値記憶領域Ｖｉ （ｉ＝１，２，…，Ｎ）をテスト及び頻度情報記憶手段１２に準備する（ステップＫ１１，Ｋ１２）。ここで、質問候補としては、回答がＹＥＳまたはＮＯであるものならなんでもよく、例えば、前述のAlmuallimらによる質問群を考えればよい。各質問Ｕｉ の評価値を計算して、評価値記憶領域Ｖｉ （ｉ＝１，２，…，Ｎ）に格納する（ステップＫ１３）。ここで、評価値は、決定木生成に採用される任意の評価関数により定められるもので、例えば、前述のエントロピーにより計算される。

【００２３】最適な質問番号格納領域Ｊとカウンターｊを１で初期化する（ステップＫ１４）。カウンターｊがＮ－１を超えないか否かをチェックし（ステップＫ１５）、超えない場合に限り、カウンターｊをカウントアップする（ステップＫ１６）。Ｖｊ がＶＪ を超える場合か、またはＶｊ がＶＪ に等しくかつＵｊ でカバーされる事例数がＵｊ でカバーされる事例数を超える場合に限り、最適な質問番号格納領域Ｊをｊで更新する（ステップＫ１７，Ｋ１８）。カウンターｊがＮ－１を超えるまで、ステップＫ１６，Ｋ１７，Ｋ１８の処理を繰り返す（ステップＫ１５）。このようにして選ばれた番号Ｊに対応する質問を、木構造を有する属性Ａに対する質問とする。

6

【００２４】図２を参照し、テスト生成手段１１の具体的処理内容を説明する。対象が、４つの属性「背丈」、「髪の色」、「目の色」、「好きな形」により、２つのクラスに対応づけられていて、「好きな形」が図５の木構造Ｔをもっている次の事例集合を考える。

Ｓ１ ＝　低い、ブロンド、青、三角形；＋
低い、黒、青、六角形；－
高い、黒、茶、凹；－
高い、ブロンド、茶、正方形；－
高い、黒、青、凹；－
低い、ブロンド、茶、凹；－
高い、赤、青、三角形；＋
高い、ブロンド、青、正方形；＋

【００２５】この例の場合、木構造をもつ属性は、「好きな形」だけなので、テスト生成手段の処理対象は、「好きな形」だけである。質問候補としてAlmuallim らによる質問群を採用すれば、ターゲット属性「好きな形」に対する質問候補は、以下の６つのテスト：

（１）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「凸」の下位ノードであるか？
（２）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「正多角形」の下位ノードであるか？
（３）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「三角形」の下位ノードであるか？
（４）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「六角形」の下位ノードであるか？
（５）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「正方形」の下位ノードであるか？
（６）訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「凹」の下位ノードであるか？

になる。従って、この場合、質問候補数Ｎは６で、以下、各質問を順にＵｉ （ｉ＝１，２，…，６）とする（ステップＫ１１）。最適な質問番号格納領域Ｊとカウンターｊを準備し、各質問候補Ｕｉ の評価値を格納する評価値記憶領域Ｖｉ （ｉ＝１，２，…，６）を準備する（ステップＫ１２）。ここで、評価値が、例えば、前述のエントロピーにより計算されるとすれば、各質問Ｕｉ の評価値を計算すると、Ｖ１ ＝０．０６１５，Ｖ２ ＝０．０６１５，Ｖ３ ＝０．２４４２，Ｖ４＝０．０１６９，Ｖ５ ＝０．０３６９，Ｖ６ ＝０．０６１５となるので、これらを順に評価値記憶領域Ｖｉ （ｉ＝１，２，…，６）に格納する（ステップＫ１３）。

【００２６】最適な質問番号格納領域Ｊとカウンターｊを１で初期化する（ステップＫ１４）。まず、カウンターｊが５を超えないか否かをチェックすると（ステップＫ１５）、超えていないのでカウンターｊをカウントアップする（ステップＫ１６）。結果として、ｊは２となる。Ｖ２ とＶ１ を比べると、共に０．０６１５で等しく、Ｕ１ でカバーされる事例は、

7

低い、ブロンド、青、三角形；＋
低い、黒、青、六角形；－
高い、ブロンド、茶、正方形；－
高い、赤、青、三角形；＋
高い、ブロンド、青、正方形；＋
の５つであり、U2 でカバーされる事例も、
低い、ブロンド、青、三角形；＋
低い、黒、青、六角形；－
高い、ブロンド、茶、正方形；－
高い、赤、青、三角形；＋
高い、ブロンド、青、正方形；＋
の５つであるから、Ｊは変化せず、１のままである（ステップＫ１７）。ｊの値は２で、５を超えないので、引き続き、ステップＫ１６，Ｋ１７，Ｋ１８を処理する。カウンターｊをカウントアップする（ステップＫ１６）。結果として、ｊは３となる。V3 とV1 を比べると、V3 が大きい（ステップＫ１７）。従って、Ｊは３で、更新される（ステップＫ１８）。ｊの値は３で、５を超えないので、引き続き、ステップＫ１６，Ｋ１７，Ｋ１８を処理する。カウンターｊをカウントアップする（ステップＫ１６）。結果として、ｊは４となる。V4 とV3 を比べると、V3 が大きい（ステップＫ１７）。

【００２７】従って、Ｊを更新せずに、ステップＫ１５に進む。ｊの値は４で、５を超えないので、引き続き、ステップＫ１６，Ｋ１７，Ｋ１８を処理する。カウンターｊをカウントアップする（ステップＫ１６）。結果として、ｊは５となる。V5 とV3 を比べると、V3 が大きい（ステップＫ１７）。従って、Ｊを更新せずに、ステップＫ１５に進む。ｊの値は５で、５を超えないので、引き続き、ステップＫ１６，Ｋ１７，Ｋ１８を処理する。カウンターｊをカウントアップする（ステップＫ１６）。結果として、ｊは６となる。V6 とV3 を比べると、V3 が大きい（ステップＫ１７）。従って、Ｊを*

8

*更新せずに、ステップＫ１５に進む。ｊの値は６で、５を超えるので、図２に示す一連の処理が終了する。このようにして、選ばれた番号Ｊは、３なので、質問U3 “訓練事例の属性値「好きな形」の属性値は、木構造Ｔ上でノード「三角形」の下位ノードであるか”が木構造を有する属性「好きな形」に対する質問として選択される。

【００２８】

【発明の効果】以上説明したように、本発明によれば、訓練事例の表現に利用する属性の少なくとも１つが木構造を有する時、木構造を有する各属性に対して最適な質問が複数あれば、その質問によってカバーされる訓練事例が一番多い質問を該属性に対する質問として選択するので、生成される決定木が木構造Ｔを有する属性で表現される場合には、該木構造Ｔの上でより上位のノードに対する質問が選択され、結果として、決定木が簡潔に表現される。

【図面の簡単な説明】

【図１】本発明の一実施形態に係る決定木生成法を実施する装置の構成を示すブロック図である。

【図２】図１に示す実施形態の作用を示すフローチャートである。

【図３】事例集合の木構造の一例を示す図である。

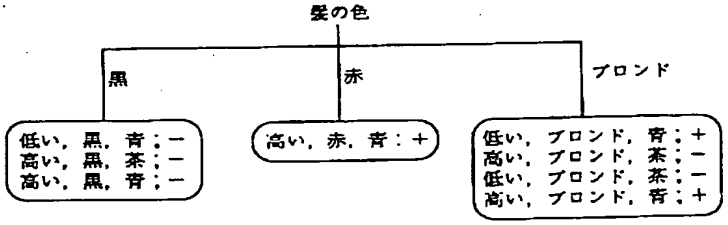
【図４】図３に示す事例集合の木構造の最終的な決定木を示す図である。

【図５】事例を表現する属性のもつ木構造の一例を示す図である。

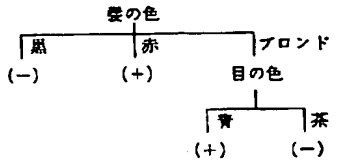
【符号の説明】

１０　事例集合族記憶手段
１１　テスト生成手段
１２　テスト及び頻度情報記憶手段
１３　分割実行手段

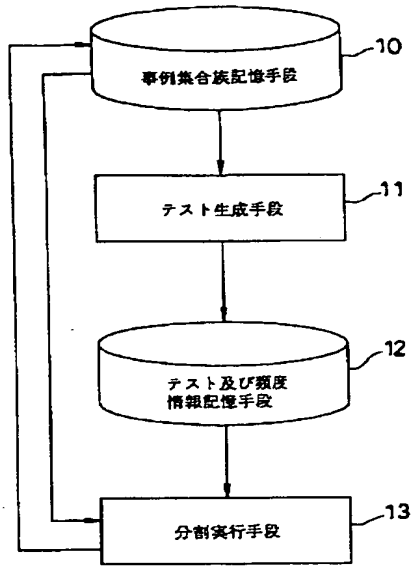
【図３】



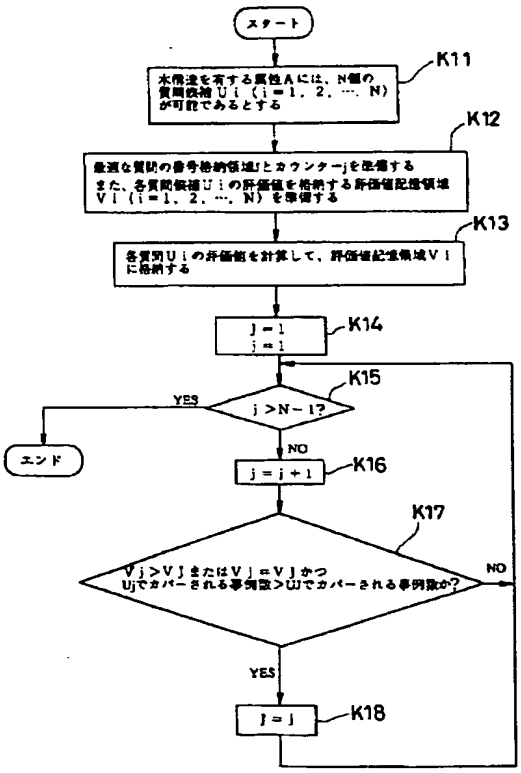
【図４】



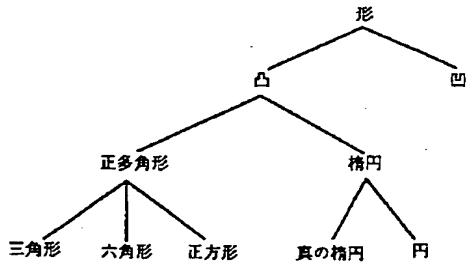
【図１】


事例集合族記憶手段
10
テスト生成手段
11
テスト及び難度
情報記憶手段
12
分割実行手段
13

【図２】


スタート
K11
K12
K13
K14
K15
YES
NO
エンド
K16
K17
NO
YES
K18

【図５】


形
凸
凹
正多角形
楕円
三角形
六角形
正方形
真の楕円
円